Video genre categorization and representation using audio-visual information

Bogdan Ionescu, Klaus Seyerlehner, Christoph Rasche, Constantin Vertan, Patrick Lambert

Video Genre Categorization and Representation using Audio-Visual Information

Bogdan IONESCU[a,c], Klaus SEYERLEHNER[b], Christoph RASCHE[a],
Constantin VERTAN[a], Patrick LAMBERT[c]
[a] LAPI-ETTI, University "Politehnica" of Bucharest, 061071, Romania
{bionescu, rasche, cvertan}@alpha.imag.pub.ro.
[b] DCP, Johannes Kepler University, A-4040 Austria
klaus.seyerlehner@jku.at.
[c] LISTIC, Polytech Annecy-Chambery, University of Savoie, 74944 France
patrick.lambert@univ-savoie.fr.

February 26, 2012

Submitted to SPIE International Journal on Electronic Imaging

ABSTRACT

We propose an audio-visual approach to video genre classification using content descriptors that exploit audio, color, temporal, and contour information. Audio information is extracted at block-level, which has the advantage of capturing local temporal information. At the temporal structure level, we consider action content in relation to human perception. Color perception is quantified using statistics of color distribution, elementary hues, color properties, and relationships between colors. Further, we compute statistics of contour geometry and relationships. The main contribution of our work lies in harnessing the descriptive power of the combination of these descriptors in genre classification. Validation was carried out on over 91 hours of video footage encompassing 7 common video genres, yielding average precision and recall ratios of $87\%-100\%$ and $77\%-100\%$, respectively, and an overall average correct classification of up to 97%. Also, experimental comparison as part of

the MediaEval 2011 benchmarking campaign demonstrated the superiority of the proposed audio-visual descriptors over other existing approaches. Finally, we discuss a 3D video browsing platform that displays movies using feature-based coordinates and thus regroups them according to genre.

Keywords: video genre classification, block-level audio descriptors, action segmentation, color perception, statistics of contour geometry, video indexing.

1 Introduction

Automatic labeling of video footage according to genre is a common requirement in indexing large and heterogeneous collections of video material. This task may be addressed, either *globally* or *locally*. Global-level approaches aim to classify videos into one of several main genres, for instance, cartoons, music, news, sports, documentaries or into more fine-grained sub-genres, for instance specific types of sports (football, hockey, etc.) and movies (drama, thriller, etc.). Local-level approaches label video segments according to specific human-centered concepts such as outdoor vs. indoor scenes, action segments and scenes showing violence (see TRECVid campaign [1]).

In this paper, we address the global classification task. Since it is related to data mining, video genre classification involves two steps: *feature extraction* and *classification*. Feature extraction and selection is one of the main critical steps determining the success of the classification task. In the literature, many sources of information have been exploited [3].

One of the least common approaches (in the context of image processing) is the use of *text-based* information, mainly due to its limited availability with video information. Text is retrieved either from scene text (e.g., graphic text, sub-titles), from the transcripts of dialogues obtained with speech recognition techniques, or from other external sources such as synopses and user tags. Bag-of-Words model approaches [4] are very common in genre classification. *Audio-based* information is more widely available than text and derived from both time and frequency domains. Common *time-domain* approaches use the Root Mean Square of signal energy (RMS) [5], sub-band informa-

tion [7], Zero-Crossing Rate (ZCR) [9], or silence ratio. *Frequency-domain* features include energy distribution, frequency centroid [9], bandwidth, pitch [10], and Mel-Frequency Cepstral Coefficients (MFCC) [8].

Of course, most video genre classification approaches rely on *visual elements*. They exploit both static and dynamic visual information in the spatial domain, for instance using color, temporal structure, objects, or motion or in the compressed domain, for instance, using MPEG coefficients [3]. *Color information* is generally derived at image level and quantified with color histograms or other low-level parameters such as predominant color, color entropy, and variance (various color spaces are used, including RGB - Red Green Blue, HSV - Hue Saturation Value, or YCbCr - Luminance, Chrominance) [11] [12] [13]. *Temporal-structure-based* information exploits temporal segmentation. A video sequence is composed of several video shots which are connected by video transitions, which can be sharp (cuts) or gradual (such as fades, dissolves) [14]. Existing approaches basically exploit their frequency of occurrence in the movie. Although some approaches use this information directly [15] (e.g., rhythm, average shot length), others derive features related to visual activity, for instance, by defining the concept of action (a high frequency of shot changes is usually related to action content) [16] [17] [18]. *Object-based features* in genre classification are generally limited to characterizing the occurrence of face and text regions in frames [22] [15] [18]. *Motion-based information* is derived either by motion detection techniques (i.e., foreground detection) or by motion estimation (i.e., prediction of pixel displacement vectors between frames, see [23]). Common features describe motion density, camera movement (global movement) and object trajectory [16] [24] [25]. Finally, less common are features computed directly in the *compressed video domain*, for example, using DCT coefficients (Discrete Cosine Transform) or embedded motion vectors from the MPEG stream [26]. Their main advantage is their immediate availability with the video file.

The efficiency of these sources in genre classification was discussed in [3]. All sources of information have advantages and disadvantages, but some prove to be more convenient than others. *Text-based* information may produce high error rates (due to automatic recognition) and is usually

computationally expensive to process (e.g., Optical Character Recognition - OCR); *object-based* information, although also computationally expensive to obtain tends to be semi-automatic (requires human confirmation); *motion information* is available in high quantities during the entire sequence (object/camera), but is insufficient by itself to distinguish between some genres, for instance, movies, sports, music. In contrast, *audio-based* information provides good discrimination and requires fewer computational resources to be obtained and processed; *color information* is not only simple to extract and inexpensive to process, but also powerful in distinguishing cinematic principles; *temporal-based* information is a popular choice and proves to be powerful as long as efficient video transition detection algorithms are employed.

The remainder of this paper is organized as follows: Section 2 discusses several genre classification approaches and situates our work accordingly. Section 3 deals with extraction of features: audio, temporal structure, color, and contour-based. Experimental results are presented in Section 4, and Section 5 concludes the paper and discusses future work.

2 Related work

The most reliable video genre classification approaches, which also target a wider range of genres, are *multi-modal*, that is, multi-source. In this section, we discuss the performance of several approaches - from single-modal (which are limited to copeing with a reduced number of genres) to multi-modal (which are able to perform more complex classifications) - we consider relevant for the present work.

A simple, single-modal approach was proposed in [24]. It addresses genre classification using only video dynamics, namely background camera motion and object motion. A single feature vector in the DCT-transformed space ensures low-pass filtering, orthogonality, and a reduced feature dimension. A classifier based on a Gaussian Mixture Model (GMM) is then used to identify three common genres: sports, cartoons, and news. Despite the limited content information used, applying the GMM model to a reduced number of genres achieves detection errors below 6%. The approach presented in [18] uses spatio-temporal information: average shot length, cut percentage, average

color difference, and camera motion (temporal) and face frames ratio, average brightness, and color entropy (spatial). Genre classification is addressed at different levels, according to a hierarchical ontology of video genres. Several classification schemes (Decision Trees and several SVM approaches) are used to classify video footage into main genres (movie, commercial, news, music, and sports) and further into sub-genres (movies into action, comedy, horror, and cartoons and sports into baseball, football, volleyball, tennis, basketball, and soccer). The highest precision achieved in video genre categorization is around 88.6% and in sub-genre categorization 97% and 81.3% for sports and movies, respectively.

However, truly multi-modal approaches also include audio information. For instance, the approach in [19] combines synchronized audio (14 Mel-Frequency Cepstral Coefficients - MFCC) and visual features (mean and standard deviation of motion vectors, MPEG-7 visual descriptors). Dimensionality of the feature vectors is reduced by means of Principal Component Analysis, and videos are classified with a GMM-based classifier. Tested with five common video genres, namely sports, cartoons, news, commercials, and music, this approach yields an average correct classification of up to 86.5%. Another example is the approach proposed in [28]. Features are extracted from four information sources: visual-perceptual information (color, texture, and motion), structural information (shot length, shot distribution, shot rhythm, shot clusters duration, and saturation), cognitive information (e.g., numbers, positions, and dimensions of faces), and aural information (transcribed text, sound characteristics). These features are used to train a parallel Neural Network system, which achieves an accuracy of up to 95% in distinguishing between seven video genres and sub-genres, namely football, cartoons, music, weather forecast, newscast, talk shows, and commercials. A generic approach to video categorization was discussed in [21]. Each video document is modeled by a Temporal Relation Matrix (TRM) which describes the relationship between video segments, that is temporal intervals related to the occurrence of a specific type of event. Events are defined based on the specificity of video features such as speech, music, applause, and speaker (audio) and color, texture, activity rate, face detection, and costume (visual). TRMs provide a similarity measure between documents. Experimental tests with several classification approaches

(mostly tree-based) and the six video genres: news, soccer, TV series, documentary, TV games, and movies yield individual genre F_{score} ratios ranging from 40% to 100% (e.g., for Random Forest with cross-validation). Another interesting approach to multimedia categorization is the cross-media retrieval method proposed in [20]. It is founded on the construction of a Multimedia Correlation Space (MMCS) which exploits semantic correlations between different multimedia modalities based on content description and co-occurrence information. The proposed video descriptors are related to color and texture (color histogram, color moment, color coherence, tamura statistics, MSRSAR texture) and aural information (RMS energy, Spectral Flux, Rolloff, Centroid). Tested with 500 multimedia objects, the system achieved correct classification rates of up to 90% in cross-media categorization.

Our approach exploits both audio and visual modalities for genre classification. Use has previously been made of these sources of information, but we approach computing these features in a novel way. The proposed *audio features* are block-level-based and have the advantage of capturing local temporal information by analyzing sequences of consecutive frames in a time-frequency representation. *Visual information* is described using temporal information, color, and structural properties. Temporal descriptors are derived using a classic confirmed approach, that is, analysis of the shot change frequency [18] [28]. However, we use a novel way of measuring action content that assesses action perception. Color information is extracted globally. In contrast to existing approaches, which mainly use local or low-level descriptors such as predominant color, color variance, color entropy, and frame based histograms [12] [18], our approach analyzes color perception. Using a color naming system, we quantify color perception with statistics of color distribution, elementary hues distribution, color properties (e.g., percentage of light colors, cold colors, saturated colors), and relationships between colors [31]. The final type of visual descriptor is related to contour information, which has rarely been exploited in genre classification [3]. Unlike most existing approaches, which describe closed region shapes (e.g., with MPEG-7 visual descriptors [27] [3]) we break contours down into segments and describe curve contour geometry both individually and relative to neighbor contours.

The principal contribution of our work, however, lies in realizing the descriptive power of the combination of these descriptors in genre classification. Extensive experimental tests conducted over 91 hours of video footage spanning seven common video genres yielded excellent classification ratios. Also, experimental comparison as part of the MediaEval 2011 benchmarking campaign proved the superiority of the proposed audio-visual descriptors compared to other approaches. Further, we tested our descriptors within a practical application, namely automatic genre categorization of video documents for potential use with media platforms (e.g., video rental, selling). We propose a prototype 3D browsing environment in which movies are displayed according to descriptor-based coordinates. Preliminary results show that movies tend to regroup according to similarities in content and genre, which is a very interesting result.

3 Content descriptors

We approach video genre categorization by exploiting audio and visual (temporal, color, and contour-based) video modalities. Our selection is motivated by the specificity of these information sources with respect to video genre.

For instance, most common video genres have very specific audio signatures, for instance, music clips contain music, news contain many monologues/dialogues, documentaries have a mixture of natural sounds, speech, and ambient music, in sports there is crowd noise, and so on. To address these particularities, we use audio descriptors related to rhythm, timbre, onset strength, noisiness, and vocal aspects.

In terms of visual information, we first extract temporal information by assessing action content and video transitions. This information is related to genre-specific cinematic principles. For instance, commercials and music clips tend to have a high visual tempo, commercials use many gradual transitions, documentaries have reduced action content, movies use gradual transitions, and so on (see also the examples in Subsection 4.1.1). The second type of visual information is related to color since different genres have different global color signatures. For example, animated

movies use specific color palettes and color contrasts (light-dark, cold-warm), music videos and movies tend to have darker colors (mainly due to the use of special effects), and sports usually show a predominant hue, such as green for soccer, white for ice hockey (see also the examples in Subsection 4.1.1). The final type of visual information is related to contours, and consequently to object shape. Different objects have different types of contours, for instance, animal silhouettes tend to be undulating with low edginess while natural scenes often contain "wiggly" and irregular shapes. Video genres tend, therefore, to have specific contour signatures: in documentaries, skyline contours dominate; in news, human faces and silhouettes are common; movies use combinations of contour shapes, such as: silhouettes, buildings, and skylines; commercials often use psycho-visual techniques (involving basic shapes such as lines, circles). Below we describe each descriptor category in more detail.

3.1 Audio features

The proposed set of audio descriptors, called *block-level audio features*, has the key advantage of capturing temporal information from the audio track at a local level.

In contrast to standard spectral audio features (e.g., Mel Frequency Spectral Coefficient, Spectral Centroid, or Spectral Roll Off), which are typically extracted from each spectral frame (capturing a time span of 20 ms) of the time-frequency representation of an audio signal, the proposed features are computed from a sequence of consecutive spectral frames called a *block*. Depending on the extracted block-level feature, a block consists of 10 up to about 512 consecutive spectral frames. Thus, local features can themselves capture temporal properties (e.g., rhythmic aspects) of an audio track over a time span ranging from half a second up to 12 seconds of audio. Blocks are analyzed at a constant rate and their frames overlap by default by 50%, which results in one local feature vector per block. These local vectors are then summarized by computing simple statistics (e.g., mean, variance, or median) separately for each dimension of the local feature vectors. A schematic diagram of this procedure is depicted in Figure 1.

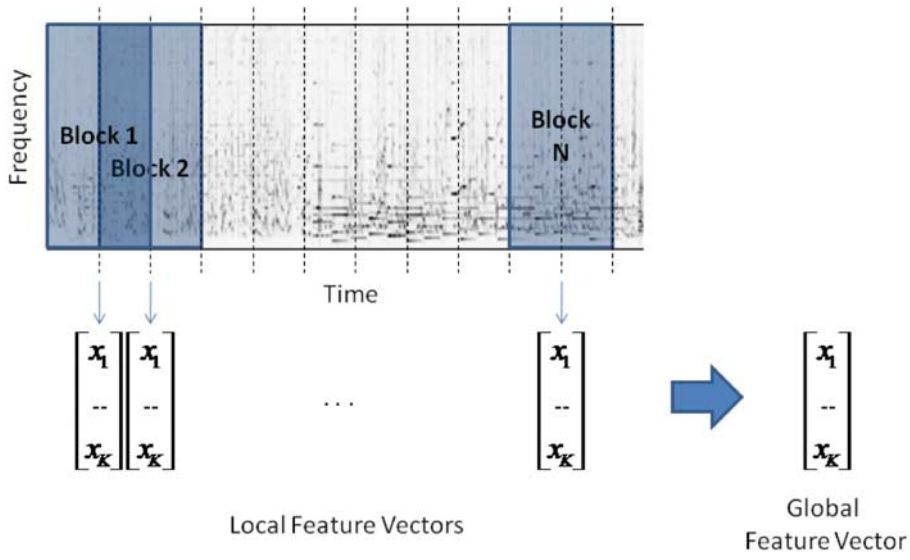

Figure 1: Processing a time (OX axis) - frequency (OY axis) representation in terms of spectral blocks (N is the number of blocks).

To obtain a perceptual time-frequency representation of the video soundtrack, the audio track is first converted into a $22kHz$ mono signal. Then we compute the short-time Fourier transform and perform a mapping of the frequency axis according to the logarithmic cent-scale because human frequency perception is logarithmic. The resulting time-frequency representation consists of 97 logarithmically spaced frequency bands. The following complex block-level audio features are derived:

Spectral pattern (SP, number of frames constituting a block: $BW = 10$, 0.9 percentile statistics): characterizes the timbre of the soundtrack by modeling the frequency components that are simultaneously active. The dynamic aspects of the signal are retained by sorting each frequency band of the block along the time axis. The block width varies depending on the extracted patterns, which allows capturing temporal information over different time spans.

Delta spectral pattern (DSP, $BW = 14$, 0.9 percentile statistics): captures the strength of onsets. To emphasize onsets, we first compute the difference between the original spectrum and a copy of the original spectrum delayed by three frames. As with the spectral pattern, each frequency band is then sorted along the time axis.

Variance delta spectral pattern ($VDSP$, $BW = 14$, variance statistics): is basically an extension of the delta spectral pattern and captures the variation of the onset strength over time.

Logarithmic fluctuation pattern (LFP, $BW = 512$, 0.6 percentile statistics): captures the rhythmic aspects of the audio signal. In order to extract the amplitude modulations from the temporal envelope in each band, periodicities are detected by computing the FFT (Fast Fourier Transform) along each frequency band of a block. The periodicity dimension is then reduced from 256 to 37 logarithmically spaced periodicity bins.

Spectral contrast pattern (SCP, $BW = 40$, 0.1 percentile statistics): roughly estimates the *"tone-ness"* of an audio track. For each frame within a block, the difference between spectral peaks and valleys in 20 sub-bands is computed, and the resulting spectral contrast values are sorted along the time axis in each frequency band.

Correlation pattern (CP, $BW = 256$, 0.5 percentile statistics). To capture the temporal relation of loudness changes over different frequency bands, the correlation coefficients for all possible pairs of frequency bands within a block are used. The resulting correlation matrix forms the correlation pattern. The correlation coefficients are computed for a reduced frequency resolution of 52 bands.

Figure 2 shows an example in which audio features were extracted from both a documentary and a music video. A typical characteristic of music signals is the presence of strong beats, which is reflected by the strong peaks in the LFP of the music video. In contrast, the flat structure of the LFP of the documentary indicates that there are no repeated percussive elements in the audio stream.

These audio features in combination with a Support Vector Machine (SVM) classifier form a highly efficient automatic music classification system. At the 2010 Music Information Retrieval Evaluation eXchange (MIREX) audio benchmark, this approach ranked first at automatic music genre classification [30]. However, the proposed approach has not yet been applied to automatic video genre classification.

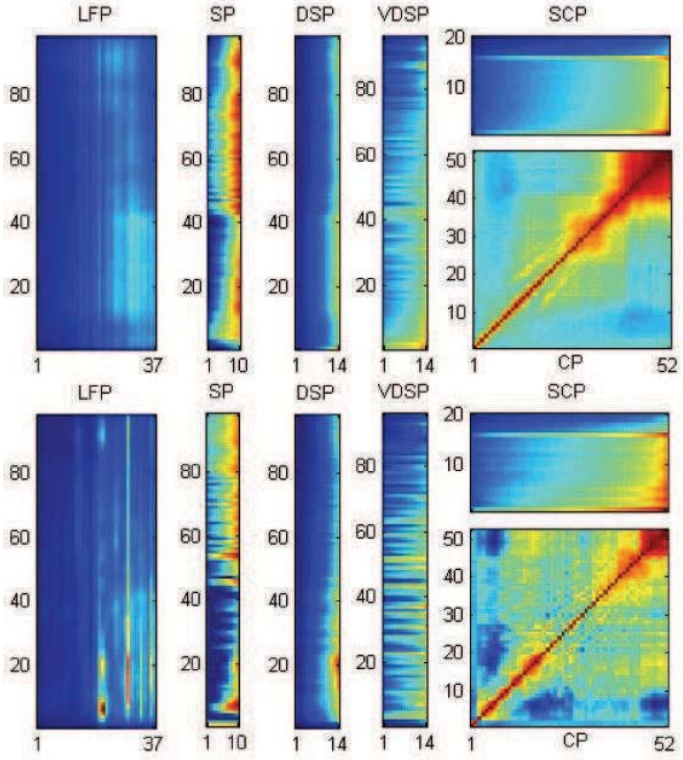

Figure 2: Visualization of the audio features of a documentary (upper plots) and a music video (lower plots). Except for the CP the OY axis of all patterns is frequency. The SCP has a lower frequency resolution. The OX axis for the SP, DSP, VDSP, SCP is related to the evolution over time. For the LFP the oX axis is periodicity. The CP is a correlation matrix.

Below we provide a preliminary test and compare our descriptors with standard MFCC in video genre classification. We tested a standard Bag-of-Frames (BoF) approach [48] in which each audio track is modeled as a single multi-dimensional Gaussian distribution over the local MFCC vectors. Two different classifiers were evaluated: the Nearest Neighbor (k-NN) classifier, to indicate retrieval performance, and a Support Vector Machine, to indicate the achievable classification performance. For nearest-neighbor classification, the distance between two Gaussian models was estimated using the KL-divergence, which is given by:

$$D_{KL}(P||Q) = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)} \, dx \tag{1}$$

where P and Q are both probability distributions. For the block-level features, the Manhattan distance was used.

Table 1 compares the classification accuracy (i.e., the number of sequences which were correctly

labeled) obtained with a Single Gaussian model over MFCC (denoted SG) to the proposed block-level feature set (denoted BLF) for the two classifiers. For our experiments we used the data set that is introduced in Section 4.

Table 1: Classification accuracy of BLF vs. MFCC.

classifier	**SG**	**BLF**
5-Nearest Neighbors	70.29%	**90.71%**
Support Vector Machines	84.10%	**91.71%**

The results clearly indicate the superiority of the block-level features over MFCC, as the improvement in classification ranges from 8% to 20%. This is also in line with the results obtained at the MediaEval 2011 Evaluation Campaign presented in Section 4.3. Table 4 shows that the video genre classification system that is based only on the proposed feature set achieved a Mean Average Performance of 10.29%. Interestingly, this approach not only outperformed an approach based on standard audio features (see team KIT), but also some systems based on textual and visual modalities. Thus, we can conclude that the proposed block-level features are more powerful in terms of music genre classification [30] and also provide an adequate audio representation for video genre classification [2]. In Section 4 we investigate whether this approach in combination with visual descriptors can help to further improve the quality of video genre classification.

3.2 Temporal descriptors

Temporal descriptors are derived by means of a classic confirmed approach, that is, analysis of the shot change frequency [18]. Unlike existing approaches, we determine the action content based on human perception.

A correct temporal description requires accurate temporal segmentation. First, we detect video transitions [14]: cuts and two of the most frequent gradual transitions - fades and dissolves. Cut detection is performed using an adaptation of the histogram-based approach proposed in [33], while fades and dissolves are detected by means of a pixel-level statistical approach [34] and analysis of

fading-in and fading-out pixels [35], respectively. The temporal descriptors are then computed as follows:

Rhythm. To capture the movie's tempo of visual change, we define a basic indicator, denoted $\zeta_T(i)$, which represents the relative number of shot changes occurring within the time interval T, starting at the frame at time index i ($T = 5$ s, determined experimentally). Based on ζ_T, we define the movie rhythm as the movie's average shot change speed, denoted $\bar{v}_T$, which is the average number of shot changes in the time interval T for the entire movie, thus $E\{\zeta_T\}$.

Action. We aim to define two opposite situations: video segments with a high action content (denoted "hot action", e.g., fast changes, fast motion, visual effects) and video segments with low action content (denoted "low action", containing mainly static scenes).

First, at a coarse level, we identify segments containing a high number of shot changes ($\zeta_T > 3.1$), which are "hot action" candidates, and a reduced number of shot changes ($\zeta_T < 0.6$), which are low action candidates (see step a in Figure 3). Thresholds were determined experimentally based on human perception: several persons were asked to manually classify video segments into the two categories mentioned. On the basis of this ground truth, we determined average ζ_T intervals for each type of action content.

To avoid over-segmentation, we merge neighboring action segments at a time distance smaller than T seconds (the size of the time window, see step b in Figure 3). Further, we remove unnoticeable and irrelevant action segments by erasing small action clips of length less than the analysis time window T. Finally, all hot action clips containing fewer than $N_s = 4$ video shots are removed because they are very likely the result of false detection and contain one or several gradual transitions (e.g., a "fade-out" - "fade-in" sequence, see step c in Figure 3). The entire process is illustrated in Figure 3. Using this information, we quantify the action content by two parameters, hot-action ratio (HA) and low-action ratio (LA):

$$HA = \frac{T_{HA}}{T_{total}}, \quad LA = \frac{T_{LA}}{T_{total}} \tag{2}$$

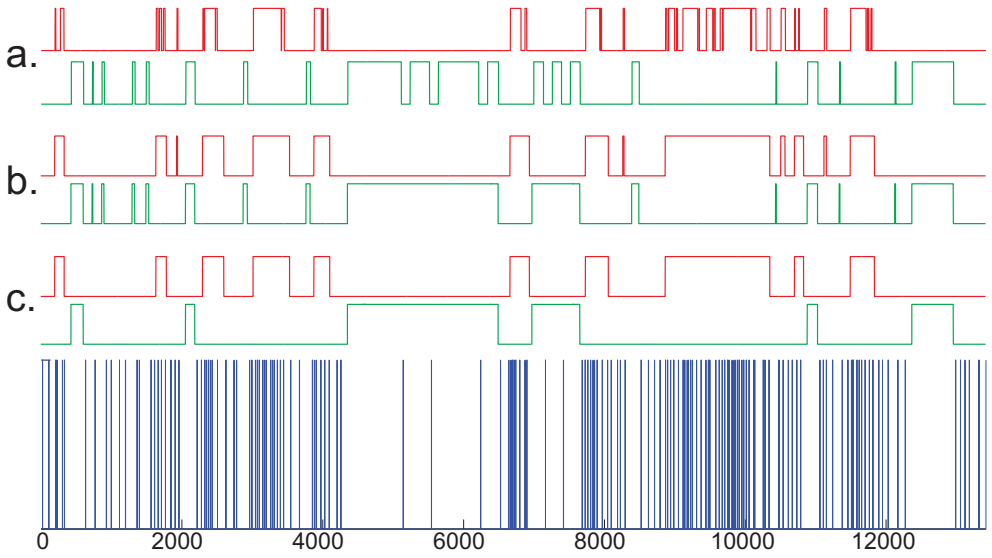

Figure 3: Action-based temporal segmentation (the OX axis is the temporal axis, vertical blue lines correspond to shot changes). "Hot action" and "low action" segments are indicated in red and green respectively. Letters denote the processing steps as described in the text.

where T_{HA} and T_{LA} represent the total lengths of hot and low action segments, respectively, and T_{total} is the total length of the movie.

Gradual transition ratio. The gradual transition ratio (GT) is computed by:

$$GT = \frac{T_{dissolves} + T_{fade-in} + T_{fade-out}}{T_{total}} \tag{3}$$

where T_x represents the total duration of all the gradual transitions of type x. This provides information about editing techniques which are specific to certain genres, such as live action feature films and artistic animated movies.

3.3 Color content

Color information is a powerful means of describing visual perception. Most existing color-based genre classification approaches are limited to using intensity-based parameters or generic low-level color features [3] such as average color differences, average brightness, average color entropy [18], variance of pixel intensity, standard deviation of gray level histograms, percentage of pixels with saturation above a given threshold [16], lighting key (measures how well light is distributed) [6],

object color, and texture.

We propose a more sophisticated strategy which addresses the perception of color content. A simple and efficient way to accomplish this is using color names; associating names with colors allows creating a mental image of a given color or color mixture. We project colors onto a color naming system [37], and color properties are described using statistics of color distribution, elementary hue distribution, color visual properties (e.g., percentage of light colors, warm colors, saturated colors, etc.), and relationships between colors (adjacency and complementarity). Color descriptors are extracted globally taking the temporal dimension into account.

Prior to parameter extraction, we project colors onto a more manageable color palette (initial images are true color). We selected the non-dithering 216 color Webmaster palette because of the high color diversity and its efficient color naming system: each color is named according to the degree of hue, saturation, and intensity [38]. Color mapping is performed with a classic dithering scheme [40], and colors are selected in the L*a*b* color space [41]. Further, the proposed color parameters are computed as follows:

Global weighted color histogram captures the global color distribution of the movie. It is computed as the weighted sum of each individual shot average color histogram:

$$h_{GW}(c) = \sum_{i=0}^{M} \left[\frac{1}{N_i} \sum_{j=0}^{N_i} h^j_{shot_i}(c) \right] \cdot \frac{T_{shot_i}}{T_{total}} \tag{4}$$

where M is the total number of video shots, N_i is the total number of frames retained from shot i (to reduce computational load, each shot is summarized by retaining $p = 10\%$ of its frames), $h^j_{shot_i}()$ is the color histogram of frame j from shot i, c is a color index from the Webmaster palette, and T_{shot_i} is the total length of shot i. The longer the shot, the more important the contribution of its histogram to the global histogram of the movie. Thus, values of $h_{GW}()$ describe global percentages of colors appearing in the movie (values are normalized to 1, i.e., a frequency of occurrence of 100%).

Elementary color histogram. This feature is computed by:

$$h_E(c_e) = \sum_{c=0}^{215} h_{GW}(c)|_{Name(c_e) \subset Name(c)} \tag{5}$$

where c_e is an elementary color from the Webmaster color dictionary, $c_e \in \Gamma_e$ with Γ_e = {"Orange", "Red", "Pink", "Magenta", "Violet", "Blue", "Azure", "Cyan", "Teal", "Green", "Spring", "Yellow", "Gray", "White", "Black"}, and $Name()$ is an operator that returns a color name from the palette dictionary.

Thus, each available color is projected in $h_E()$ onto its elementary hue, while saturation and intensity information are disregarded. This mechanism removes susceptibility to color fluctuations (e.g., illumination changes) and provides information about predominant hues.

Color properties. These parameters aim to describe color perception by means of light/dark, saturated/non-saturated, warm/cold colors and color richness by quantifying color variation/diversity. Using the previously determined histogram information in conjunction with the color naming dictionary, we define several color ratios. For instance, the light color ratio, P_{light}, which reflects the percentage of bright colors in the movie, is computed by:

$$P_{light} = \sum_{c=0}^{215} h_{GW}(c)|_{W_{light} \subset Name(c)} \tag{6}$$

where c is the index of a color whose name (provided by $Name(c)$) contains one of the words defining brightness, and $W_{light} \in$ {"light", "pale", "white"}. Using the same reasoning and keywords specific to each color property, we define:

- dark color ratio, denoted P_{dark}, where $W_{dark} \in$ {"dark", "obscure", "black"};
- hard color ratio, denoted P_{hard}, which reflects the number of saturated colors. $W_{hard} \in$ {"hard", "faded"} $\cup \Gamma_e$, where Γ_e is the elementary color set (see eq. 5, elementary colors are 100% saturated colors);
- weak color ratio, denoted P_{weak}, which is opposite to P_{hard}, $W_{weak} \in$ {"weak", "dull"};
- warm color ratio, denoted P_{warm}, which reflects the number of warm colors; in art, some hues are commonly perceived to exhibit levels of warmth, namely: "Yellow", "Orange", "Red", "Yellow-Orange", "Red-Orange", "Red-Violet", "Magenta", "Pink" and "Spring";

- cold color ratio, denoted P_{cold}, where "Green", "Blue", "Violet", "Yellow-Green", "Blue-Green", "Blue-Violet", "Teal", "Cyan" and "Azure" describe coldness.

Further, we capture movie color richness with two parameters. Color variation, P_{var}, which represents the number of significant colors, is defined as:

$$P_{var} = \frac{Card\{c|h_{GW}(c) > \tau_{var}\}}{216} \tag{7}$$

where c is a color index, h_{GW} is the global weighted histogram defined in eq. 4, and $Card()$ is the cardinal function, which returns the size of a data set. We consider a color significant if it has a frequency of occurrence in a movie of more than 1% (i.e., $\tau_{var} = 0.01$). Color diversity, P_{div}, which reflects the number of significant color hues in the movie, is defined using the same principle, but based on the elementary color histogram h_E.

Color relationship. The final two parameters are related to the concept of perceptual relationships between colors in terms of adjacency and complementarity. The parameter, P_{adj} reflects the number of perceptually similar colors in the movie (neighborhood pairs of colors on a perceptual color wheel, e.g., Itten's color wheel [42] [38]), and P_{compl} reflects the number of perceptually opposite color pairs (antipodal).

3.4 Structural content

The final set of parameters provides information based on structure, that is, on contours and their relations. So far, contour information has been exploited to a very limited extent in genre classification. For instance, some approaches use MPEG-7-inspired contour descriptors [3] [27], such as texture orientation histograms, edge direction histograms, edge direction coherence, [43], which do not exploit real contour geometry and properties.

Our approach, in contrast, proposes a novel method which uses curve partitioning and curve description [29]. The contour description is based on a characterization of geometric attributes of each individual contour, for instance, degree of curvature, angularity, and "wiggliness". These

attributes are used as parameters in a high-dimensional image vector and have been exploited successfully in a (statistical) classification task. For instance, the system achieved the benchmark for urban and natural scene collection in [44], and ranked among the upper third of all performances in the photo-annotation task of the ImageCLEF competition [45].

Contour characterization. Contour processing starts with edge detection, which is performed with the Canny edge detection algorithm [32]. For each contour, a type of curvature space is created. This space is then abstracted into spectra-like functions, from which a number of geometric attributes, such as the degree of curvature, angularity, circularity, symmetry and "wiggliness", are derived. In addition to these geometric parameters, a number of "appearance" parameters are extracted. They are based on simple statistics obtained from the luminance values extracted along the contour, such as contrast (mean and standard deviation, abbreviated c_m, and c_s respectively) and "fuzziness", obtained by convolution of the image with a blob filter (f_m, and f_s, respectively). In summary, for a given image with n extracted and partitioned contours, we obtain a list of 7 geometric and 4 appearance attributes for each contour. For each attribute, a 10-bin histogram with n values, is generated.

Pair relations. In addition to individual contour attributes, we also obtain attributes for pairs of contours. Contour segments are first grouped as a list of all $n!$ pairs. From this long list of pairs, only a subset (approximately $3 \times n$) is selected based on spatial proximity, meaning that their contour endpoints are either proximal or in the proximity of other segments.

For each selected pair, a number of geometric attributes is determined: the angular direction of the pair (γ_p), the distance between the proximal contour endpoints (d_c), the distance between the distal contour end points (d_o), the distance between segment center (middle) points (d_m), the average segment length (l), the symmetry of the two segments (y), the degree of bendiness of each segment (b_1 and b_2), the structural biases ($\hat{s}$) which express to what degree the pair alignment is an L feature ($\hat{s}_L$), a T feature ($\hat{s}_T$), or a "closed" feature (two curved segments facing each other as '( )', $\hat{s}_{()}$). In total, 12 geometric relational attributes are extracted for the selected pairs. Again,

for each attribute a 10-bin histogram is generated.

Figure 4 shows an example of the representative power of these descriptors in image-based categorization. We present the results obtained by similarity-based contour search in the Corel collection (60000 images) using three concepts: "landscape", "entrance", and "people". The contour of the first image in each row is the selected sample contour, the remaining images in each row contain the most similar contours. Regular objects can be associated with particular salient contour signatures (see the blue contours in Figure 4) and retrieved accordingly with good recognition rates [29]. This information can be very useful in tackling our genre classification problem.

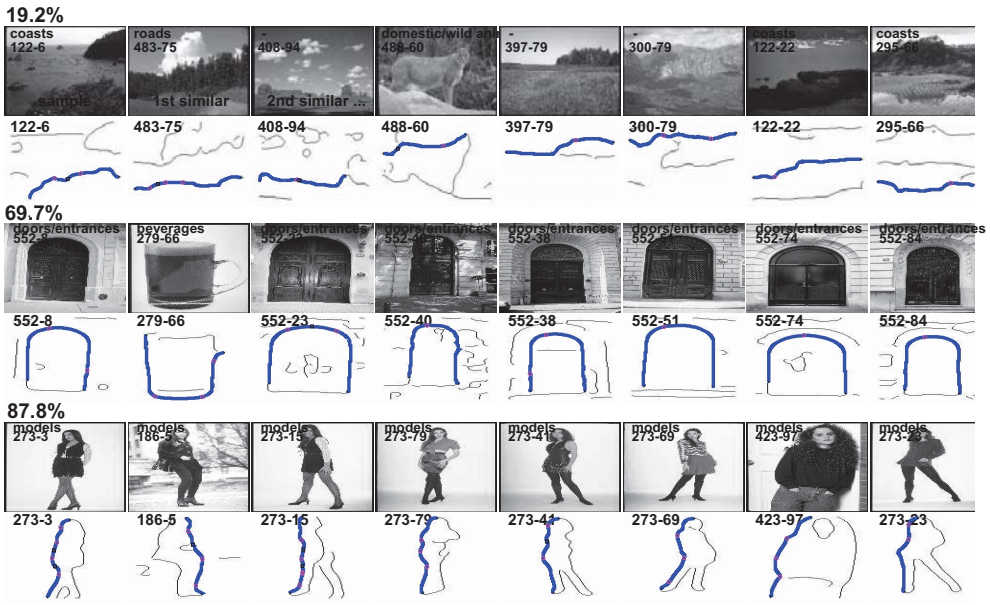

Figure 4: Similarity-based contour search for all contours of the entire Corel collection (60000 images). The contour of the first image in each row is the selected sample contour, the remaining images in each row contain the most similar contours. The percentage on the left denotes correct basic-level categorization for the first 99 similar images.

The structural information is extracted only from a summary of the movie. In this case, we retained around 100 images that are evenly distributed with respect to video transitions. As previously mentioned, at image level, contour properties are captured with histograms. To address the temporal dimension - at sequence level - the resulting concatenated feature vectors are averaged to form so the structure signature of the movie (see also the examples in Figure 7).

4 Experimental results

In this section, we present several experimental scenarios to validate the proposed descriptors: the example descriptors emphasize their specificity with respect to video genre, classification tests demonstrate their power for genre classification, and 3D feature-based representation show their potential in real-world browsing applications. Finally, we also present a comparative benchmark evaluation.

4.1 Classification experiments

To asses the representative power of the proposed content descriptors, we conducted several experiments. For validation, we selected seven of the most common video genres, namely *animated movies*, *commercials*, *documentaries*, *movies*, *music videos*, *news broadcast*, and *sports*. Classifying these genres is challenging due to content similarity: animated movies include natural scenes (we used not only cartoons but also artistic movies, see [39]), commercials also include cartoons, news include scenes from sports and scenes resembling documentaries, music clips tend to have visual patterns similar to those of commercials, and so on.

The data set comprises over 91 hours of video footage (30 sequences per genre). Video material was retrieved from several TV programmes:

- animated movies: 1230 min, long, short clips and series, sources: Folimage - France, Disney, Pixar, and DreamWorks animation companies;
- commercials: 15 min, sources: TV commercials from the 1980s and David Lynch clips;
- documentaries: 1320 min, on the topics wildlife, ocean, cities and history, sources: BBC, IMAX, and Discovery Channel;
- movies: 1317 min, long, episodes and sitcom, e.g., Friends, X-Files, Sex and the City series;
- music: 150 min, pop, rock and dance video clips, source: MTV Channel;

- news broad cast: 1320 min, source: TVR Romanian National Television Channel;
- sports, 115 min, various short clips from the Internet.

Several supervised strategies were used for classification. As the choice of training data may distort the accuracy of the results, we used a cross-validation approach: for each experiment we tested all possible combinations of training and test data. Additionally, we varied the amount of training data (percentage split from 10% to 70%) and tested different combinations of descriptors.

To assess performance at genre level, we evaluated average precision (P) and recall (R) ratio:

$$P = \frac{\overline{TP}}{\overline{TP} + \overline{FP}}, \quad R = \frac{\overline{TP}}{\overline{TP} + \overline{FN}} \tag{8}$$

where $\overline{TP}$, $\overline{FP}$, and $\overline{FN}$ are the *average* numbers of correct detections (True Positives), false detections (False Positives) and non-detections (False Negatives), respectively. Averaging was performed over all repetitions for a given amount of training data.

As a global measure of performance, we computed F_{score} ratio and average correct classification ($\overline{CD}$):

$$F_{score} = 2 \cdot \frac{P \cdot R}{P + R}, \quad \overline{CD} = \frac{\overline{N_{GD}}}{N_{total}} \tag{9}$$

where $\overline{N_{GD}}$ is the average number of good classifications, and N_{total} is the number of test sequences.

4.1.1 Descriptor examples

For a preliminary analysis of the representative power of the proposed descriptors with respect to video genre we show the average audio, color-action, and contour descriptors in Figures 5, 6 and 7, respectively, for each of the seven genres. In general, each genre behaves differently.

The most visible differences can be found in the audio descriptors, where measures such as logarithmic fluctuation pattern, spectral pattern, and delta spectral pattern discriminate well between all genres. In contrast, color-action and contour descriptors tend to emphasize the specificity of some genres. For instance, commercials and music clips have a high visual rhythm and action

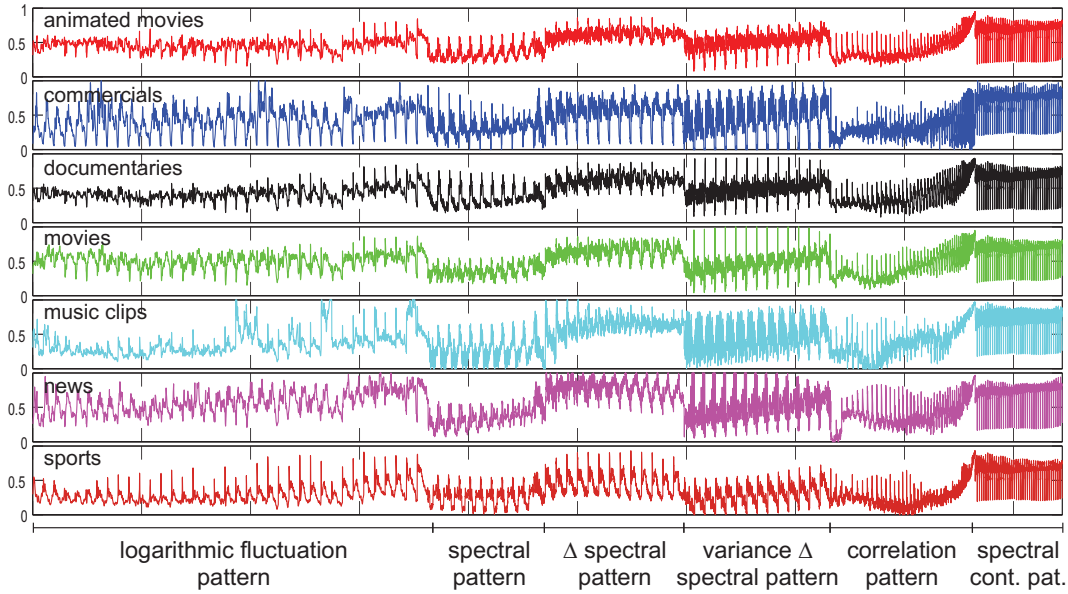

Figure 5: Average audio feature vectors for each genre ("cont." stands for contrast, and "pat." for pattern, see Section 3.1).

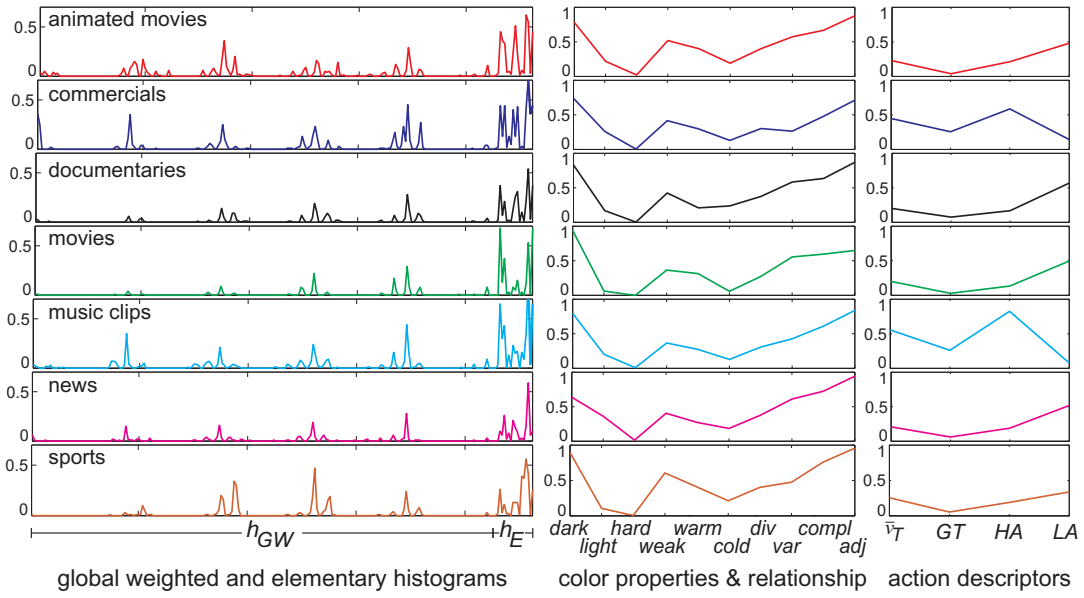

Figure 6: Average color-action feature vectors for each genre (see Section 3.2 and 3.3).

content (see $\bar{v}_T$ and HA in Figure 6); animated movies have a different color pattern (more colors are used, see h_{GW}) and most of the hues are used in significant amounts (see h_E); movies and documentaries tend to have reduced action content; sports scenes have a predominant hue (see the high peak in h_E); commercials show an important symmetry of contours (see high values in contour relationship in Figure 7). The representative power of the proposed features is further corroborated

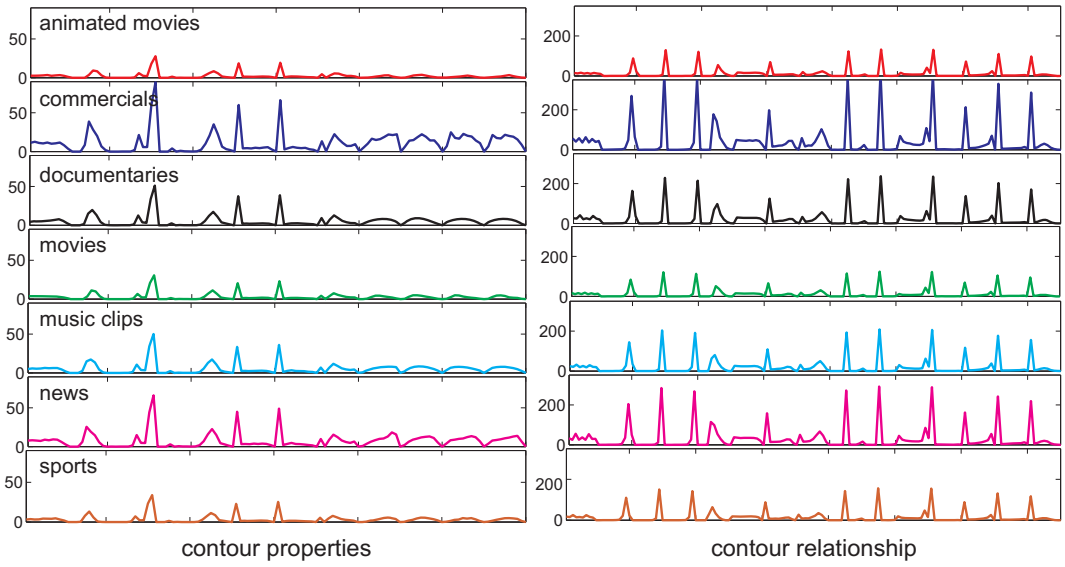

Figure 7: Average contour feature vectors for each genre (see Section 3.4).

by the following classification experiments.

4.1.2 One-genre-at-a-time binary classification

The first classification experiment consisted of retrieving one genre at a time. To this end, we selected three binary classifiers, namely K-Nearest Neighbors (KNN, with k=1, cosine distance, and majority rule), Support Vector Machines (SVM, with a linear kernel) and Linear Discriminant Analysis (using Principal Component Analysis to reduce dimensionality). Method parameters were tuned to appropriate values based on preliminary experiments.

Figure 8 plots average precision against recall (see eq. 8) for different amounts of training data and different descriptor combinations. The results are very promising considering the diversity of sub-genres within each genre (see video sources at the beginning of Section 4). The best classification results we obtained were $P \in [87.5\%; 100\%]$ ($P > 95\%$ for music, news, commercials, and sports), and $R \in [77.6\%; 100\%]$ ($R > 95\%$ with animated movies and commercials excluded).

Figure 9 presents overall F_{score} and correct detection $\overline{CD}$ for all genres (which takes into account correct classification in both classes, i.e., target genre and others, see eq. 9). The highest F_{score} of 90.6% was obtained using 70% of the data for training, while the overall correct classification ratio

ranges from 92.2% to 97.2%. The overall performance is very good, even for a reduced amount of training data, as $F_{score} > 83\%$ for 20% training data, and $\overline{CD} > 92\%$ for as little as 10% training data.

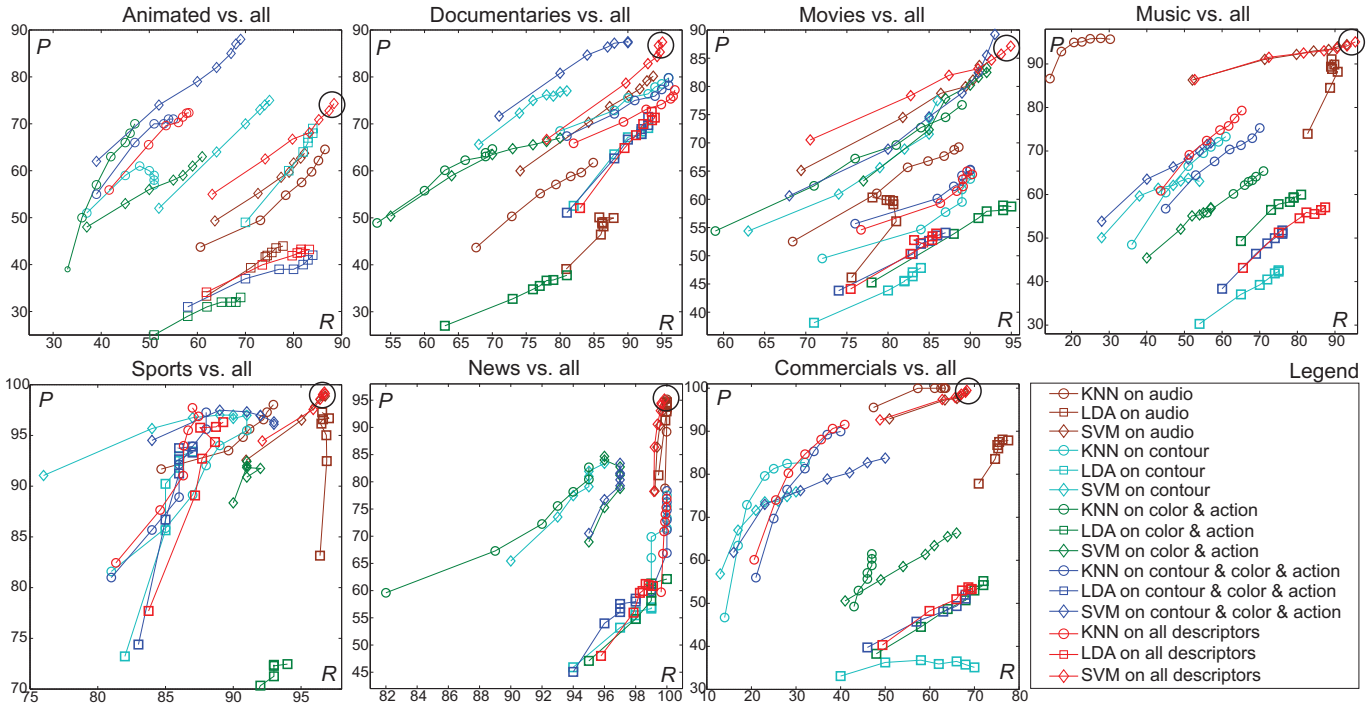

Figure 8: Precision (P) against recall (R) for different runs and amounts of training data (increases along the curves from 10% to 70%; the encircled results are detailed in Table 2).

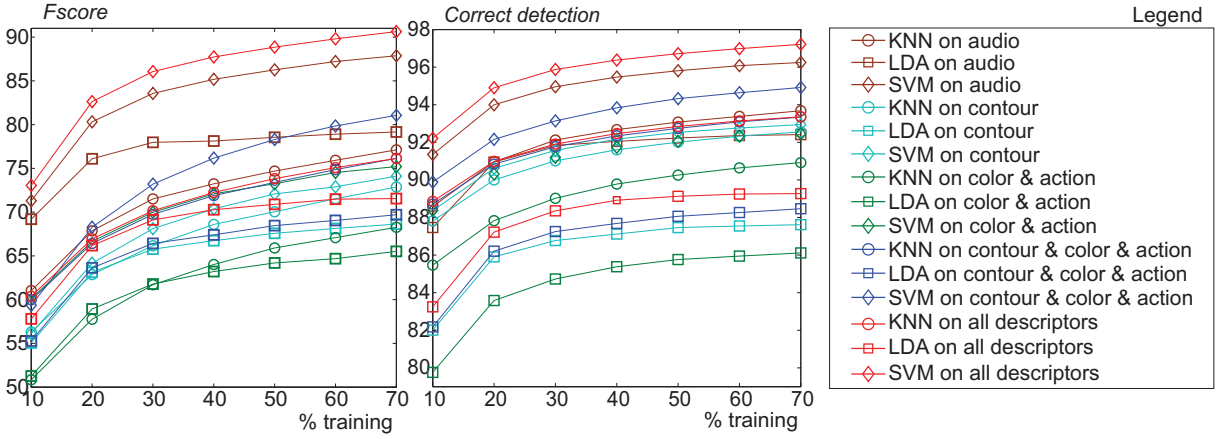

Figure 9: Overall average F_{score} and correct classification ($\overline{CD}$) for all genres against the amount of training data.

The most interesting result, however, is that each descriptor set harnesses different properties of the video content, as the most efficient approach (both in terms of overall classification performance

and genre precision and recall) is the combination of all audio-visual descriptors (see the SVM results in Figure 8). Table 2 summarizes the precision and recall ratios for this case ("docum." stands for documentaries and "comm." for commercials; these results are marked with black circles in Figure 8).

Table 2: SVM vs. KNN and LDA (using all audio-visual descriptors).

genre	Precision (P)			Recall (R)		
	SVM	KNN	LDA	SVM	KNN	LDA
animated	**74.3%**	72.3%	43.2%	**88.4%**	58.2%	83.3%
docum.	**87.4%**	77.2%	72.6%	**95.1%**	96.3%	93.5%
movies	**87.1%**	65%	53.9%	**94.9%**	89.6%	85.8%
music	**95.1%**	79.3%	57%	**95.4%**	65.2%	87.3%
sports	**99.3%**	97.7%	96.3%	**96.7%**	86.9%	89.2%
news	**95.2%**	76.9%	60.8%	**99.8%**	99.9%	99.1%
comm.	**99.5%**	91.5%	53.3%	**68.3%**	40.9%	69.4%

Globally, the lowest accuracy is obtained for animated movies and commercials, mainly because their content is heterogeneous and resembles that of other genres. For instance many commercials include animation, music clips are similar to commercials, movies may contain commercial-like contents, etc. The best performance (as anticipated) was obtained for genres with a certain repetitiveness in content structure, that is, news and sports (average precision or recall up to 100%).

From the angle of the information sources, audio information proves to be highly efficient compared to visual information, and leads to very good classification ratios (see Figure 8). At genre level, audio features are more accurate at retrieving music, sports, news, and commercials, as these genres have specific audio patterns. Using contour and color-action information alone proves to be less efficient. Compared to color-action parameters, contour parameters yield better performance for documentaries, sports, and news, which have salient contour signatures such as skyline contours and silhouettes of people (see Figure 8). Compared to contour parameters, color-action features perform better for music, commercials, movies, and news (which can be attributed to the specific rhythm and color diversity, see also Figure 8). Compared to audio descriptors, visual descriptors used in combination are more discriminative for animated movies, movies, and documentaries. Finally, the best performance in classifying each individual genre was achieved by using all audio-visual

information available.

4.1.3 Multi-class classification

In the final classification test, all genres were to be classified simultaneously. We used all audio-visual descriptors and the multi-class SVM classifier proposed in [47] (i.e., the descriptor-method combination which previously provided the best classification results). Since we used the same classification strategy as described at the beginning of Section 4, we varied the amount of training data (from 10% to 70%) and classification was repeated for all possible combinations of training and test data.

Figure 10 plots the average precision against recall, average F_{score} and average correct classification ($\overline{CD}$) ratios we obtained. Discriminative power is maintained even when addressing all genres at once. The best classification results are summarized in Table 3. At genre level, we achieved $P \in [78\%; 94\%]$ and $R \in [62\%; 100\%]$. Compared to the binary one-genre-at-a-time classification approach (see Subsection 4.1.2), precision and recall tend to be slightly lower for some of the genres (see also Table 2), but still significant (on average above 85%) compared to the results described in the literature (see Section 2).

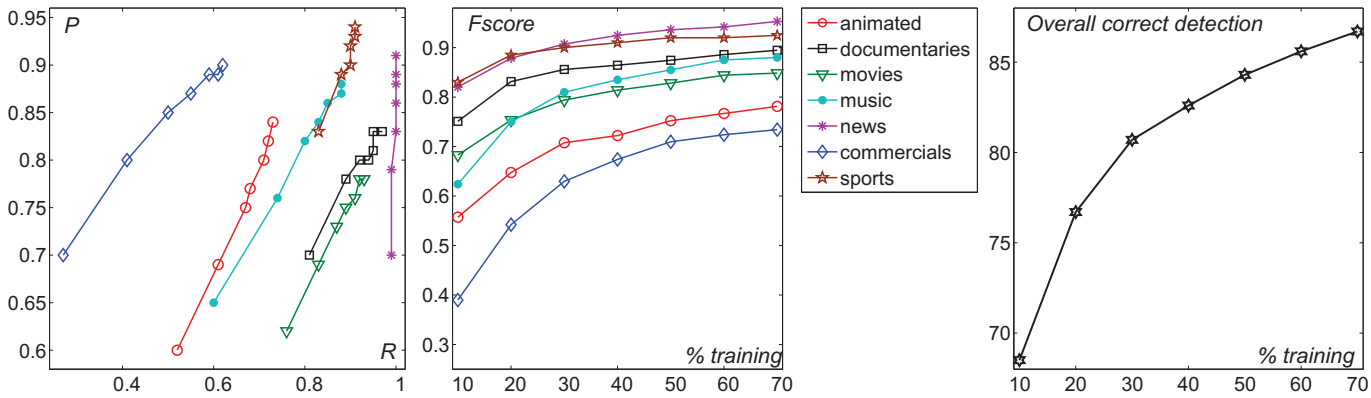

Figure 10: SVM multi-class classification results, from left to right: average precision vs. recall, overall average F_{score} and correct classification ($\overline{CD}$) for all genres against the amount of training data used.

Similar to previous results, the best classification performance was obtained for genres with some

Table 3: SVM multi-class results.

genre	Precision (P)	Recall (R)	F_{score}
news	**91%**	**100%**	95.3%
sports	**94%**	**91%**	92.5%
documentaries	83%	**97%**	89.5%
music	88%	88%	88%
movies	78%	**93%**	84.8%
animated	84%	73%	78%
commercials	**90%**	62%	73.4%

degree of repetitiveness and specificity in content structure: news ($F_{score} = 95.3\%$) followed closely by sports ($F_{score} = 92.5\%$), and then documentaries ($F_{score} = 89.5\%$) and music ($F_{score} = 88\%$). For news and sports we achieved F_{score} ratios above 80% even when only 10% of the data was used for training (see Figure 10). Classification performance was less satisfactory for genres with heterogeneous content and content similar to that of other genres, that is commercials ($F_{score} = 73.4\%$) and animated movies ($F_{score} = 78\%$).

In terms of overall average correct classification ($\overline{CD}$), we achieved $\overline{CD} > 80\%$ when using just 30% of the data for training, which means that out of 147 test sequences, 119 sequences were correctly labeled. The accuracy increases with the amount of training data, for example, for 50% training data, the $\overline{CD} = 84.3\%$; thus, out of 105 test sequences, 89 sequences were correctly labeled. The highest correct classification was 86.7%.

Figure 11 plots the average confusion matrix obtained for 50% training data. The genres that are most often mislabeled are animated movies and commercials, followed by music. With increasing the amounts of training data, confusion proportions tend to remain similar to those presented (for reasons of brevity we do not present all confusion matrices). Nevertheless, classification errors remain greatly reduced (few sequences for each genre - confusion matrix diagonal values are significant with respect to others).

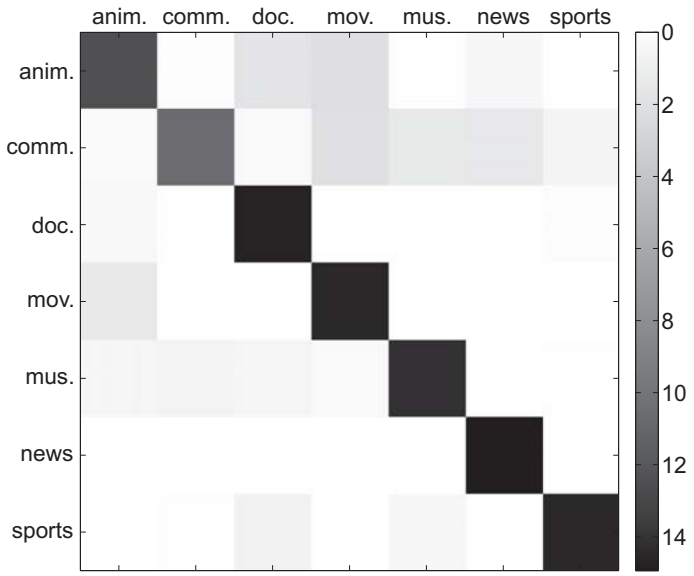

Figure 11: Average confusion matrix (50% training data, i.e., 105 sequences, 15 per genre; abbreviations: "anim." - animated, "comm." - commercials, "doc." - documentaries, "mov." - movies, "mus." - music).

4.2 Content-based representation

The following experiment was conducted at application level. We sought to simulate a video browsing environment in which sequences were to be represented using the proposed descriptors. We have developed a client-server architecture which provides a virtual 3D browsing environment for video databases [46]. Movies are displayed in a spherical coordinate system, and each movie is represented by one key frame. The user interface resembles that of Google Earth (by which we were inspired): the user flies virtually through "constellations of movies".

For displaying movies, we combined all descriptors, since this provided the best classification results. As we are restricted to only 3 axes in selecting the most representative components, we used Principal Component decomposition. Each movie is displayed according to:

- inclination (denoted θ) - the first PCA component (normalized to $[0; \pi]$),
- azimuth (φ) - second PCA component (normalized to $[0; 2\pi]$),
- radius (r) - third PCA component (normalized to $[0; 1]$).

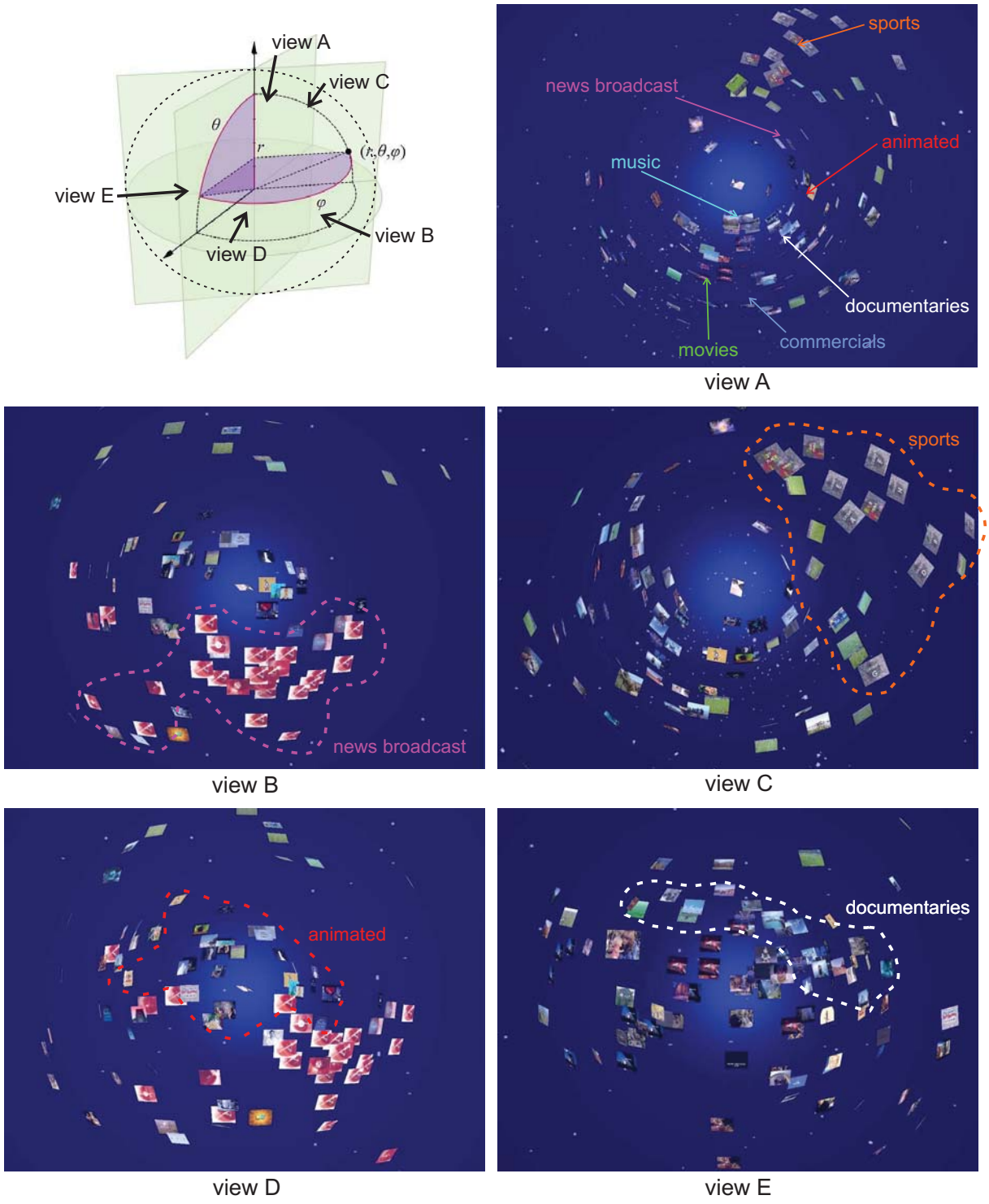

Figure 12: Feature-based 3D movie representation in a spherical coordinate system (inclination-θ, azimuth-φ, radius-r). Each movie from the data set is represented by a point with which we associate an image vignette. Views A to E are screenshots taken from different perspectives (the points of view used are shown in the chart). In views A-E, representative genres are annotated (a demo is available at `http://imag.pub.ro/~bionescu/index_files/MovieGlobe.avi`).

Several screenshots taken from different angles are presented in Figure 12. Interestingly, although we use only the first three principal components (which account for up to 94% of the initial data variance), movies from particular genres form clusters. Due to similarity in content and structure, the most clearly grouped genres are news (see view B in Figure 12) and sports (see view C in Figure 12). Other genres tend to be more "interleaved", as one might expect, since even human observers find it difficult to draw sharp distinctions between genres (see also the observations at the beginning of Section 4). Nevertheless, sequences of the same genre tend to regroup around a basis partition (see the examples in Figure 12, e.g., animated movies - view D, documentaries - view E).

Coupled with genre labeling provided by a classification mechanism (e.g., SVM), this could be a powerful genre-based browsing tool. However, although they illustrate the potential of our descriptors, these results are only preliminary, and more detailed tests must be conducted.

4.3 MediaEval benchmark

To provide a more standardized evaluation of the proposed descriptors in comparison with other approaches, we present the results obtained for the MediaEval 2011 (Benchmark Initiative for Multimedia Evaluation) Video Genre Tagging Task [2]. The following 26 video genres were to be tagged automatically: "*art*", "*autos and vehicles*", "*business*", "*citizen journalism*", "*comedy*", "*conferences and other events*", "*documentary*", "*educational*", "*food and drink*", "*gaming*", "*health*", "*literature*", "*movies and television*", "*music and entertainment*", "*personal of auto-biographical*", "*politics*", "*religion*", "*school and education*", "*sports*", "*technology*", "*environment*", "*mainstream media*", "*travel*", "*video blogging*", "*web development and sites*" and "*default category*" (containing movies that cannot be assigned to any one of the previous categories).

Each participant was provided with a development set consisting of 247 sequences, unequally distributed across genres (some genre categories contained very few - even just one or two - examples). This initial set served as a reference point for the development of the proposed solution. The participants were encouraged to build their own training sets if required by their approach.

Consequently, to provide a consistent training data set for classification, we extended the data set to up to 648 sequences. The final classification task was performed using a test set consisting of 1727 sequences (approximatively 350 hours of footage). After testing various machine learning techniques (we used the Weka environment, see `http://www.cs.waikato.ac.nz/ml/weka/`) on the development data, the most accurate results were achieved again with a linear SVM approach using all audio-visual descriptors. Therefore, we used this approach for the final classification run.

Performance was assessed by computing the overall Mean Average Precision (MAP) as defined by TRECVid [1] (see also trec_eval scoring tool at `http://trec.nist.gov/trec_eval/`). In Table 4 we compare our results with several other approaches using various modalities of the video - from textual (e.g., speech transcripts, user tags, metadata - provided by the organizers with the data sets) to audio-visual information[1]. A detailed overview of the results was presented in [2].

The proposed descriptors achieved an overall MAP of up to 12% (see team RAF in Table 4), which - considering difficulty of the task - is significant. Also, these were the best results obtained using audio-visual information alone. Using descriptors such as cognitive information (face statistics), temporal information (average shot duration, distribution of shot lengths) [28], audio (MFCC, zero crossing rate, signal energy), color (histograms, color moments, autocorrelogram - denoted "autocorr."), and texture (co-occurrence - denoted "co-occ.", wavelet texture grid, edge histograms) with SVM resulted in MAPs below 1% (see team KIT in Table 4); clustered SURF features in combination with SVM achieved a MAP of up to 9.4% (see team TUB in Table 4). We achieved better performance even compared to some classic text-based approaches, for instance, the Term Frequency-Inverse Document Frequency (TF-IDF, MAP 9.8%, see team UAB in Table 4) and the Bag-of-Words (MAP 5.5%, see team SINAI in Table 4) approaches. Compared to visual information, audio descriptors seem to provide better discriminative power for this task.

It must be noted, however, that the results presented in Table 4 cannot be definitive, as the classification approaches were not trained and set up strictly comparably. Teams were allowed

[1]the following notations were used: Terrier IR is an information retrieval system, see `http://terrier.org/`; Delicious is a social tagging site, see `http://del.icio.us/`.

Table 4: Comparative results: MediaEval benchmark [2] (selection).

descriptors	**modality**	**method**	**decision**	**MAP**	**team**
speech transcripts	text	Support Vector Machines	ranked list	11.79%	LIA
speech transcripts	text	Bag-of-Words + Terrier IR	ranked list	10.31%	SINAI
speech transcripts, metadata	text	Bag-of-Words + Terrier IR	ranked list	10.73%	SINAI
speech transcripts, metadata, user tags	text	Bag-of-Words + Terrier IR	ranked list	11.15%	SINAI
speech transcripts	text	Bag-of-Words	ranked list	5.47%	SINAI
speech transcripts	text	TF-IDF + cosine dist.	binary	6.21%	UAB
speech transcripts, metadata	text	TF-IDF + cosine dist.	binary	9.34%	UAB
speech transcripts, metadata, user tags	text	TF-IDF + cosine dist.	binary	9.4%	UAB
speech transcripts, Delicious tags	text	BM25F [36] + Kullback - Leibler divergence	ranked list	11.03%	UNED
speech transcripts, Delicious tags, metadata	text	BM25F [36] + Kullback - Leibler divergence	ranked list	11.11%	UNED
metadata	text	Negative multinomial divergence	ranked list	39.37%	TUD
MFCC, zero cross. rate, signal energy	audio	multiple SVMs	binary	0.1%	KIT
proposed	**audio**	**SVM with linear kernel**	**binary**	**10.29%**	RAF
clustered SURF	visual	Bag-of-Visual-Words + SVM with Radial Basis kernel	binary	9.43%	TUB
hist., moments, autocorr., co-occ., wavelet, edge hist.	visual	multiple SVMs	binary	0.35%	KIT
cognitive (face statistics [28])	visual	multiple SVMs	binary	0.1%	KIT
structural (shot statistics [28])	visual	multiple SVMs	binary	0.3%	KIT
proposed	**visual**	**SVM with linear kernel**	**binary**	**3.84%**	**RAF**
color, texture, aural, cognitive, structural	audio, visual	multiple SVMs	binary	0.23%	KIT
proposed	**audio, visual**	**SVM with linear kernel**	**binary**	**12.08%**	**RAF**
clustered SURF, metadata	visual, text	Naive Bayes, SVM + serial fusion	binary	30.33%	TUB

to access other sources of information than those proposed in the competition). For instance, we used 648 sequences for training, whereas team KIT used up to 2514 sequences. Most text-based approaches used query expansion techniques (e.g., Wordnet - see `http://wordnet.princeton.edu/`, Wikipedia - see `http://en.wikipedia.org`). Nevertheless, these results provide a good overview and (crude) comparative ranking of the performance of various methods and in consequence of the proposed descriptors.

In conclusion, the competition results show that the most efficient retrieval approach remains

the inclusion of textual information, as it provides a higher semantic level of description than audio-visual information. The average MAP obtained by including textual descriptors is around 30% (e.g., see team TUB in Table 4), which is still hard to achieve using only video information.

5 Conclusions

We have addressed global video genre categorization using four categories of content descriptors: block-level audio features, temporal-based descriptors, color perceptual descriptors and statistics of contour geometry. These sources of information have previously been exploited, but our approach provides a novel way of computing these content descriptors. The main contribution of our work, however, lies in harnessing the descriptive power of the combination of these descriptors in genre classification. We validated our approach in several experiments using over 91 hours of video footage encompassing seven common video genres (animated, movies, news, sports, commercials, movies and documentaries). Furthermore, experiments conducted at the MediaEval 2011 benchmarking campaign proved the superiority of our proposed audio-visual descriptors compared to other validated approaches.

In individual genre retrieval (binary classification), we achieved average precision and recall ratios of $87\% - 100\%$ and $77\% - 100\%$, respectively, while average correct classification was up to 97%. Using only audio information proves to be - compared to visual information - highly efficient in tackling this task. Audio features are more accurate when classifying music, sports, news, and commercials. Visual descriptors are more discriminative than audio for animations, movies, and documentaries. The best performance, however, is obtained when all audio-visual descriptors are used in combination. Retrieving all genres simultaneously (multi-class classification) produced good results: the F_{score} achieved ranged from 95.3% for news to 73% for commercials. Experimental comparison as part of the MediaEval 2011 benchmarking campaign demonstrated also the superiority of the proposed audio-visual descriptors over other existing approaches

Finally, we tested the potential of the proposed descriptors at the application level. Movies

displayed according to feature-based coordinates in a prototype 3D browsing environment tend to regroup according to similarities in content and genre. Coupled with genre labeling provided by a classification mechanism (e.g., SVM), this could be a powerful genre-based browsing tool.

Future improvements will mainly consist of approaching sub-genre categorization and consideration of the constraints of very large scale approaches (millions of sequences and tens of genre concepts).

6 Acknowledgments

This work has been supported by the Sectoral Operational Programme Human Resources Development 2007-2013 of the Romanian Ministry of Labour, Family and Social Protection through the Financial Agreement POSDRU/89/1.5/S/62557 and by the Austrian Science Fund (FWF): L511-N15.

The authors would like to thank CITIA - The City of Moving Images and Folimage Animation Company for providing them with access to their animated movie database and for their support.

We also acknowledge the 2011 Genre Tagging Task of the MediaEval Multimedia Benchmark [2] for providing the test data set.

References

[1] A. F. Smeaton, P. Over, W. Kraaij, "High-Level Feature Detection from Video in TRECVid: a 5-Year Retrospective of Achievements, Multimedia Content Analysis," Theory and Applications, Springer Verlag-Berlin, pp. 151-174, ISBN 978-0-387-76567-9, 2009.

[2] M. Larson, A. Rae, C.-H. Demarty, C. Kofler, F. Metze, R. Troncy, V. Mezaris, Gareth J.F. Jones (eds.), Working Notes Proceedings of the MediaEval 2011 Workshop at Interspeech 2011, vol. 807, CEUR-WS.org, ISSN 1613-0073, http://ceur-ws.org/Vol-807/, Pisa, Italy, September 1-2, 2011.

[3] D. Brezeale, D.J. Cook, "Automatic Video Classification: A Survey of the Literature," IEEE Trans. on Systems, Man, and Cybernetics, Part C: Applications and Reviews, 38(3), pp. 416-430, 2008.

[4] G. Forman, "An Extensive Empirical Study of Feature Selection Metrics for Text Classification," Journal of Machine Learning Research, 3, pp. 1289-1305, 2003.

[5] Z. Rasheed, M. Shah, "Movie Genre Classification by Exploiting Audio-Visual Features of Previews," IEEE Int. Conf. on Pattern Recognition, 2, pp. 1086-1089, 2002.

[6] Z. Rasheed, Y. Sheikh, M. Shah, "On the use of Computable Features for Film Classification," IEEE Trans. Circuits and Systems for Video Technology, 15, pp. 5264, 2003.

[7] M. H. Lee, S. Nepal, U. Srinivasan, "Edge-based Semantic Classification of Sports Video Sequences," IEEE Int. Conf. on Multimedia and Expo, 2, pp. 157-160, 2003.

[8] U. Srinivasan, S. Pfeiffer, S. Nepal, M. Lee, L. Gu, S. Barrass, "A Survey of Mpeg-1 Audio, Video and Semantic Analysis Techniques," Multimedia Tools and Applications, 27(1), pp. 105-141, 2005.

[9] Z. Liu, J. Huang, Y. Wang, "Classification of TV Programs based on Audio Information using Hidden Markov Model," IEEE Workshop on Multimedia Signal Processing, pp. 27-32, 1998.

[10] J. Fan, H. Luo, J. Xiao, L. Wu, "Semantic Video Classification and Feature Subset Selection under Context and Concept Uncertainty," ACM/IEEE Conference on Digital Libraries, pp. 192-201, 2004.

[11] M. S. Drew, J. Au, "Video Keyframe Production by Efficient Clustering of Compressed Chromaticity Signatures," ACM Int. Conf. on Multimedia, pp. 365-367, 2000.

[12] D. Brezeale, D.J. Cook, "Using Closed Captions and Visual Features to Classify Movies by Genre," Int. Workshop on Multimedia Data Mining, 2006.

[13] X. Gibert, H. Li, D. Doermann, "Sports Video Classification using HMMs," Int. Conf. on Multimedia and Expo, 2, pp. II-345-348, 2003.

[14] R. Lienhart, "Reliable Transition Detection in Videos: A Survey and Practitioners Guide", Int. Journal of Image and Graphics, 1(3), pp. 469-486, 2001.

[15] G. Wei, L. Agnihotri, N. Dimitrova, "TV Program Classification based on Face and Text Processing," IEEE Int. Conf. on Multimedia and Expo, 3, pp. 1345-1348, 2000.

[16] B. T. Truong, C. Dorai, S. Venkatesh, "Automatic Genre Identification for Content-Based Video Categorization," Int. Conf. on Pattern Recognition, IV, pp. 230-233, 2000.

[17] R. Jadon, S. Chaudhury, K. Biswas, "Generic Video Classification: An Evolutionary Learning based Fuzzy Theoretic Approach," Conf. on Computer Vision, Graphics, and Image Processing, 2002.

[18] X. Yuan, W. Lai, T. Mei, X.-S. Hua, X.-Q. Wu, S. Li, Automatic Video Genre Categorization using Hierarchical SVM," IEEE Int. Conf. on Image Processing, pp. 2905-2908, 2006.

[19] L.-Q. Xu, Y. Li, "Video classification using spatial-temporal features and PCA," International Conference on Multimedia and Expo, pp. 485488, 2003.

[20] Y. Yang, D. Xu, F. Nie, J. Luo, Y. Zhuang, "Ranking with Local Regression and Global Alignment for Cross Media Retrieval," ACM International Conference on Multimedia, pp. 175-184, October 1924, 2009, Beijing, China.

[21] Z. Al A. Ibrahim, I. Ferrane, P. Joly, "A Similarity-Based Approach for Audiovisual Document Classification Using Temporal Relation Analysis," EURASIP Journal on Image and Video Processing, *doi* : 10.1155/2011/537372, 2011.

[22] G. Wei, I.K. Sethi, "Face Detection for Image Annotation," Pattern Recognition Letters, 20, pp. 1313-1321, 1999.

[23] A. Bovik, "The Essential Guide to Video Processing," Academic Press, ISBN 978-0-12-374456-2, 2009.

[24] M.J. Roach, J.S.D. Mason, "Video Genre Classification using Dynamics," IEEE Int. Conf. on Acoustics, Speech and Signal Processing, pp. 1557-1560, Utah, USA, 2001.

[25] G. Y. Hong, B. Fong, A. Fong, "An Intelligent Video Categorization Engine," Kybernetes, 34(6), pp. 784-802, 2005.

[26] H. Wang, A. Divakaran, A. Vetro, S.-F. Chang, H. Sun, "Survey of Compressed-Domain Features used in Audio-Visual Indexing and Analysis," Journal of Visual Communication and Image Representation, 14(2), pp. 150-183, 2003.

[27] T. Sikora, "The MPEG-7 Visual Standard for Content Description An Overview," IEEE Trans. on Circuits and Systems for Video Technology, 11(6), pp. 696-702, 2001.

[28] M. Montagnuolo, A. Messina, "Parallel Neural Networks for Multimodal Video Genre Classification", Multimedia Tools and Applications, 41(1), pp. 125-159, 2009.

[29] C. Rasche: "An Approach to the Parameterization of Structure for Fast Categorization," Int. Journal of Computer Vision, 87(3), pp. 337-356, 2010.

[30] K. Seyerlehner, M. Schedl, T. Pohle, P. Knees, "Using Block-Level Features for Genre Classification, Tag Classification and Music Similarity Estimation," 6th Annual Music Information Retrieval Evaluation eXchange (MIREX-10), Utrecht, Netherlands, August 9-13, 2010.

[31] B. Ionescu, C. Vertan, P. Lambert, A. Benoit, "A Color-Action Perceptual Approach to the Classification of Animated Movies," ACM Int. Conf. on Multimedia Retrieval, Trento, Italy, 17-20 April, 2011.

[32] J. Canny, "A Computational Approach To Edge Detection," IEEE Trans. on Pattern Analysis and Machine Intelligence, 8(6), pp. 679-698, 1986.

[33] B. Ionescu, V. Buzuloiu, P. Lambert, D. Coquin, "Improved Cut Detection for the Segmentation of Animation Movies," IEEE Int. Conf. on Acoustic, Speech and Signal Processing, Toulouse, France, 2006.

[34] W.A.C. Fernando, C.N. Canagarajah, D.R. Bull, "Fade and Dissolve Detection in Uncompressed and Compressed Video Sequence," IEEE Int. Conf. on Image Processing, Kobe, Japan, pp. 299-303, 1999.

[35] C.-W. Su, H.-Y.M. Liao, H.-R. Tyan, K.-C. Fan, L.-H. Chen, "A Motion-Tolerant Dissolve Detection Algorithm," IEEE Trans. on Multimedia, 7(6), pp. 1106-1113, 2005.

[36] J. Pérez-Iglesias, J. R. Pérez-Agüera, V. Fresno, Y. Z. Feinstein, "Integrating the Probabilistic Models BM25/BM25F into Lucene". CoRR, abs/0911.5046, 2009.

[37] P. Kay, T. Regier, "Resolving the Question of Color Naming Universals," Proceedings of the National Academy of Sciences of the United States of America, 100(15), pp. 9085-9089, 2003.

[38] B. Ionescu, D. Coquin, P. Lambert, V. Buzuloiu: "A Fuzzy Color-Based Approach for Understanding Animated Movies Content in the Indexing Task," Eurasip Journal on Image and Video Processing, doi:10.1155/2008/849625, 2008.

[39] B. Ionescu, L. Ott, P. Lambert, D. Coquin, A. Pacureanu, V. Buzuloiu, "Tackling Action - Based Video Abstraction of Animated Movies for Video Browsing", SPIE - Journal of Electronic Imaging, Vol. 19, No. 3, 2010.

[40] R. W. Floyd and L. Steinberg, "An Adaptive Algorithm for Spatial Gray Scale", Proc. Int. Symp. Digest of Technical Papers, pp. 3637, 1975.

[41] W. K. Pratt, "Digital Image Processing," John Wiley & Sons, Hoboken, NJ, USA, 2007.

[42] J. Itten, "The Art of Color: The Subjective Experience and Objective Rationale of Color," Reinhold, New York, NY, USA, 1961.

[43] A. Hauptmann, R. Yan, Y. Qi, R. Jin, M. Christel, M. Derthick, M.- Y. Chen, R. Baron, W.-H. Lin, T. D. Ng, "Video Classification and Retrieval with the Informedia Digital Video Library System," Text Retrieval Conference, 2002.

[44] A. Oliva, A. Torralba, "Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope," Journal of Computer Vision, 42(3), pp. 145-175, 2001.

[45] C. Rasche, C. Vertan, "A Novel Structural-Description Approach for Image Retrieval," CLEF Notebook Papers, `http://clef2010.org/resources/proceedings/clef2010labs_submission_56.pdf`, 2010.

[46] B. Ionescu, A. Marin, P. Lambert, D. Coquin, C. Vertan, "A Content-Driven System Architecture for Tackling Automatic Cataloging of Animated Movie Databases," Int. Journal of Digital Library Systems, 1(2), pp. 1-23, 2010.

[47] I. Tsochantaridis, T. Hofmann, T. Joachims, Y. Altun, "Support Vector Learning for Interdependent and Structured Output Spaces," Int. Conf. on Machine Learning, 2004.

[48] J.-J. Aucouturier, F. Pachet, "A Scale-Free Distribution of False Positives for a Large Class of Audio Similarity Measures," Pattern Recognition, 41(1), pp. 272-284, 2008.

Biographies

Bogdan Ionescu is currently a lecturer at University "Politehnica" of Bucharest-Romania. He holds a B.S. degree in applied electronics (2002) and an M.S. degree in computing systems (2003), both from University Politehnica of Bucharest. He also holds a Ph.D. degree in image processing (2007) from, both, the University of Savoie and University "Politehnica" of Bucharest. Between 2006 and 2007, he held a temporary Assistant Professor position at Polytech'Savoie, Uni-

versity of Savoie. His scientific interests cover video processing, video retrieval, computer vision, software engineering, and computer science. He is a Member of IEEE, SPIE, ACM, and GDR-ISIS.

Klaus Seyerlehner is a postdoctoral researcher at the Department of Computational Perception at Johannes Kepler University in Linz, Austria. He holds an M.S. (2006) and a Ph.D. (2011) degree in computer science, both from Johannes Kepler University. His main research interests cover the fields of digital music signal processing, pattern recognition, machine learning, statistics and recommender systems.

Christoph Rasche obtained a PhD degree in Computational Neuroscience, with a focus on Neuromorphic Engineering, at ETH Zürich in 1999. He then held postdoctoral and research positions at Caltech (Pasadena), Penn-State (State College), Justus-Liebig University (Giessen, Germany) and is now at the Universitatea Politehnica din Bucuresti (Romania). He has written 17 journal articles and a book entitled "The Making of a Neuromorphic Visual System". His research interest is the field of visual recognition, specifically shape retrieval, image classification, image understanding, eye movements, and gaze-guided human-computer interaction.

Constantin Vertan holds an image processing and analysis tenure at the Image Processing and Analysis Laboratory from the Faculty of Electronics, Telecommunications and Information Technology at the "Politehnica" University of Bucharest (UPB). He was an invited professor at INSA de Rouen and the University of Poitiers (France). For his contributions in image processing

he was awarded the "In tempore opportuno" award (2002) by the UPB and the "In hoc signo vinces" award (2004) by the Romanian National Research Council. His research interests are general image processing and analysis, CBIR, fuzzy and medical image processing applications. He is a member of SPIE, senior member of IEEE and secretary of the Romanian IEEE Signal Processing Chapter.

Patrick Lambert received his engineering degree in electrical engineering in 1978, and a PhD in signal processing in 1983, both from the National Polytechnic Institute of Grenoble, France. He is currently a Full Professor at the School of Engineering of the University of Savoie, Annecy, France and a member of the Informatics, Systems, Information and Knowledge Processing Laboratory (LISTIC), Annecy, France. His research interests are in the field of image and video analysis, and he currently focuses on non-linear color filtering and video semantic indexing.